Charotar University of Science and Technology (CHARUSAT)
Faculty of Technology and Engineering
Devang Patel Institute of Advance Technology & Research (DEPSTAR)
Subject: OCAIML4001: Responsible & Safe AI Systems
PRACTIBCAL LIST

Sr no.	Practical	CO/PO
1	**Problem Definition:** **Uncovering Bias in Social Media Sentiment Analysis** An AI ethics researcher at a startup is tasked with auditing a sentiment analysis tool developed for a major social media platform. This tool is designed to automatically filter toxic comments and flag negative posts. However, complaints have arisen indicating that the system disproportionately flags content related to certain identities (e.g., comments containing gender or race-specific terms), even when such content is not offensive. The objective is to identify, analyze, and report any bias present in the system and propose improvements using fairness metrics and natural language processing techniques. **Key Questions / Analysis / Interpretation to be evaluated during/after Implementation** 1. **Model Training:** Train a sentiment classifier using the Twitter Sentiment140 dataset with Hugging Face or other NLP tools. 2. **Bias Testing:** Create and test identity-sensitive inputs to observe differences in sentiment predictions. 3. **Fairness Analysis:** Apply fairness metrics like Demographic Parity and Equal Opportunity to assess and interpret prediction bias. **Supplementary Problems –** 1. How do predictions change when gendered terms are altered? 2. What difference does balanced vs. imbalanced training data make in model fairness? **Key Skills to be addressed –** • Training and evaluation of NLP-based sentiment models • Creating fairness-sensitive test datasets • Applying group fairness metrics • Interpreting bias and communicating results **Applications** – Hate speech and toxicity detection systems, Social media content moderation **Learning Outcome –** • Detect biases in NLP models	CO1, CO2/PO1, PO2

	• Understand and apply fairness metrics • Recognize ethical risks in real-world NLP applications	
	Dataset/Test Data (Source and Description If Applicable) : **Sentiment140 (Twitter)**: Contains tweets labeled as positive or negative sentiment **Tools/Technology To Be Used**: Python, Hugging Face Transformers, Fairlearn, Scikit-learn, Pandas	
	* **Total Hours of Problem Definition Implementation** * **Total Hours of Engagement = Time required to implement solution + Modification time, if necessary + Testing of the solution by the faculty member (may be with suggested test data) :** 3 + 1 hours	
	Post Laboratory Work Description: • Submit a bias analysis report including test cases, fairness metrics, and recommendations • Reflect on societal impacts • Propose potential improvements	
	Evaluation Strategy Including Viva: • Model Training and Accuracy - 2 • Test Set Design for Bias Detection - 2 • Fairness Metric Calculation and Interpretation- 2 • Quality of Report and Visualization – 2 • Viva (Conceptual Understanding + Case - Discussion) - 2 **Total 10 Marks** **Feedback on Problem Definition Implementation (Satisfaction Level 0 to 4, where 0 is lowest,1 is poor,2 is average, 3 is good, 4 is excellent) (This can asked for group of practical belongs to same tool/concept/technology):** **Satisfaction Level (0–4):** • 0: Not satisfied • 1: Poor • 2: Average • 3: Good • 4: Excellent **Feedback to be collected for:** • Ease of understanding the problem • Clarity of dataset and objectives • Relevance of tools/techniques • Difficulty level and learning impact	

2	**Problem Definition**: **Preventing Toxic Outputs in AI-Powered Language Systems.** A social media company is preparing to launch an AI-powered virtual assistant that helps users compose posts and comments. During internal testing, the assistant occasionally generates offensive, abusive, or politically sensitive content in response to seemingly neutral or provocative prompts. Concerned about user safety and brand reputation, the company assigns the AI Ethics and Safety Team to audit the model's output, detect toxic behavior, and recommend mitigation strategies before deployment. **Key Questions / Analysis / Interpretation to be evaluated during/after Implementation** 1. Generate text using GPT-2 or similar. 2. Evaluate toxicity using Google Perspective API or OpenAI moderation tool. 3. Analyze triggers for toxic output and suggest mitigation strategies. **Supplementary Problems –** 1. How does toxicity vary with different prompt phrasing (e.g., polite vs. aggressive)? 2. Can fine-tuning reduce toxic output? 3. Compare toxicity levels between open-source models and controlled commercial models. 4. Test responses in different languages and analyze cross-lingual toxicity. **Key Skills to be addressed –** - Text generation using NLP models - Toxicity detection using third-party classifiers - Prompt engineering and mitigation strategies - Analysis of model behavior and unintended bias - Ethical considerations in deploying LLMs **Applications –** Content filtering in virtual assistants, Ethical deployment of generative AI tools in public platforms **Learning Outcome –** • Understand how generative AI can produce toxic content • Learn toxicity detection techniques • Appreciate challenges in safe language model deployment	CO1, CO2/PO1, PO2
	Dataset/Test Data (Source and Description If Applicable) : Jigsaw Unintended Bias in Toxicity Classification Dataset (Kaggle)	

	Tools/Technology To Be Used: Python, Hugging Face Transformers, Perspective API/OpenAI API	
	*** Total Hours of Problem Definition Implementation** *** Total Hours of Engagement = Time required to implement solution + Modification time, if necessary + Testing of the solution by the faculty member (may be with suggested test data) :** 3 + 1 hours	
	Post Laboratory Work Description: Submit a report with: • Generated toxic and non-toxic outputs • Observations from toxicity scores • Mitigation strategies used and their impact • Ethical reflections on deploying such models	
	Evaluation Strategy Including Viva: Model Training and Accuracy - 2 Toxicity detection and analysis - 2 Implementation of mitigation techniques- 2 Quality of Report and Visualization - 2 Viva (Conceptual Understanding + Case - Discussion) - 2 **Total 10 Marks** **Feedback on Problem Definition Implementation (Satisfaction Level 0 to 4, where 0 is lowest,1 is poor,2 is average, 3 is good, 4 is excellent) (This can asked for group of practical belongs to same tool/concept/technology):** **Satisfaction Level (0–4):** • 0: Not satisfied • 1: Poor • 2: Average • 3: Good • 4: Excellent **Feedback to be collected for:** • Ease of understanding the problem • Clarity of dataset and objectives • Relevance of tools/techniques • Difficulty level and learning impact	
3	**Problem Definition**: **Defending Vision Models Against Adversarial Attacks in High Stakes Applications.**	CO3, CO4/ PO3, PO4

	A healthcare startup is deploying an AI system for automated medical image diagnosis, using deep learning to identify diseases from X-rays and CT scans. During security evaluation, it is found that small, imperceptible changes to the images can trick the model into making incorrect diagnoses, such as missing a tumor or predicting disease in a healthy image. To ensure safety and reliability, the development team is assigned to perform robustness testing using adversarial attacks and implement defense mechanisms to improve the model's resilience. **Key Questions / Analysis / Interpretation to be evaluated during/after Implementation** Apply and analyze adversarial perturbations to test the robustness of computer vision models (e.g., ResNet) to attacks. 1. Load a pre-trained image classifier (e.g., ResNet18 on CIFAR-10 or ImageNet). 2. Use adversarial attack algorithms (FGSM, PGD) to generate adversarial examples. 3. Evaluate model accuracy on clean vs. adversarial inputs. 4. Apply adversarial training or defensive distillation to increase robustness and compare improvements. **Supplementary Problems –** 1. Compare the effectiveness of different attacks (FGSM, PGD, DeepFool) on the same model 2. Apply adversarial attacks on different architectures (e.g., ResNet vs. VGG) 3. Test how adversarial noise affects explainability (using GradCAM) 4. Evaluate robustness under different training regimes (dropout, normalization, etc.) **Key Skills to be addressed –** • Implementation of adversarial attack algorithms in vision models • Defensive strategies to harden models • Visual analysis of perturbations and decision boundaries • Critical evaluation of model reliability and safety **Applications –** Medical image diagnosis (robustness in critical predictions), Surveillance and security AI, Defense against AI exploitation and vulnerabilities **Learning Outcome –** • Understand how vision models can be exploited using minimal perturbations • Analyze the effects of robustness-enhancing techniques • Learn how to quantify and improve reliability and safety of AI models under attack scenarios	
	Dataset/Test Data (Source and Description If Applicable) : CIFAR-10 or MNIST	

	Tools/Technology To Be Used: Python, PyTorch, Torchvision, Foolbox or CleverHans (for attacks), NumPy, Matplotlib	
	*** Total Hours of Problem Definition Implementation** *** Total Hours of Engagement = Time required to implement solution + Modification time, if necessary + Testing of the solution by the** **faculty member (may be with suggested test data) :** 3 + 1 hours	
	Post Laboratory Work Description: Submit a **report** including: • Sample original vs. adversarial images • Model performance metrics pre- and post- attack • Implementation of defense techniques and re-evaluation • Reflections on limitations and practical implications	
	Evaluation Strategy Including Viva: Attack Implementation and Accuracy comparison - 2 Defense Strategy Implementation - 2 Visualization and Interpretation - 2 Quality of Report and Visualization – 2 Viva (Conceptual Understanding + Case - Discussion) - 2 **Total 10 Marks** **Feedback on Problem Definition Implementation (Satisfaction Level 0 to 4, where 0 is lowest,1 is poor,2 is average, 3 is good, 4 is excellent) (This can asked for group of practical belongs to same tool/concept/technology):** **Satisfaction Level (0–4):** • 0: Not satisfied • 1: Poor • 2: Average • 3: Good • 4: Excellent **Feedback to be collected for:** • Ease of understanding the problem • Clarity of dataset and objectives • Relevance of tools/techniques Difficulty level and learning impact	

4	**Problem Definition**: **Detect hidden trojans in neural networks using trigger inputs and specialized datasets to reveal security vulnerabilities.** A defense technology company plans to deploy a pre-trained AI model for object detection in surveillance drones. This model was outsourced from an external vendor. During testing, unusual behavior is observed—certain inputs with specific patterns consistently trigger incorrect predictions, such as identifying a "stop sign" as a "green light." Suspecting a backdoor (trojan) attack, the company assigns the AI Security Team to conduct a forensic audit of the model. The goal is to simulate and detect trojans using trigger-based inputs, compare behavior between clean and trojaned models, and assess potential security vulnerabilities before deployment. **Key Questions / Analysis / Interpretation to be evaluated during/after Implementation** 1. Load and evaluate behavior of trojaned and clean models 2. Simulate trigger-based inputs and observe targeted misclassification 3. Compare performance with and without the trigger **Supplementary Problems –** 1. What happens if the trigger is partially occluded? 2. Can model pruning or fine-tuning remove the trojan? 3. Compare behavior under data poisoning vs. model poisoning. 4. Test with different trigger shapes, colors, or positions. **Key Skills to be addressed –** • Understanding and simulating trojan attacks • Evaluation of model robustness to malicious perturbations • Trigger input generation • Visualization of internal neural network behavior • Model performance analysis and threat diagnosis **Applications –** Neural network supply chain validation and model sharing, Model certification and deployment safety audits **Learning Outcome –** • Understand how trojans are embedded in neural networks • Learn to simulate and detect backdoor attacks • Gain practical experience with real-world model vulnerability scenarios	CO3, CO4/ PO3, PO4

	Dataset/Test Data (Source and Description If Applicable) : Trojan Detection Challenge Dataset (NeurIPS)	
	Tools/Technology To Be Used: Python, PyTorch / TensorFlow, Torchvision, NumPy, Matplotlib	
	*** Total Hours of Problem Definition Implementation** *** Total Hours of Engagement = Time required to implement solution + Modification time, if necessary + Testing of the solution by the faculty member (may be with suggested test data) :** 3 + 1 hours	
	Post Laboratory Work Description: Submit a **report** that includes: • Clean vs. trojaned model comparison • Sample trigger images and observed misclassifications • Performance metrics and analysis • Reflections on how such vulnerabilities affect real-world AI systems	
	Evaluation Strategy Including Viva: • Model Behavior Analysis (Clean vs. Trojaned) - 2 • Trigger Generation and Misclassification Observations - 2 • Clarity and Insight in Report - 2 • Security Risk Interpretation - 2 • Viva (Conceptual Understanding + Case - Discussion) - 2 **Total 10 Marks** **Feedback on Problem Definition Implementation (Satisfaction Level 0 to 4, where 0 is lowest,1 is poor,2 is average, 3 is good, 4 is excellent) (This can asked for group of practical belongs to same tool/concept/technology):** **Satisfaction Level (0–4):** • 0: Not satisfied • 1: Poor • 2: Average	

	• 3: Good • 4: Excellent **Feedback to be collected for:** • Ease of understanding the problem • Clarity of dataset and objectives • Relevance of tools/techniques • Difficulty level and learning impact	
5	**Problem Definition**: **Model Explainability with LIME and SHAP on Text Classification** A fintech company has developed a sentiment analysis system to evaluate customer feedback and automatically escalate negative reviews for priority handling. However, the customer support team raises concerns that some negative reviews are being misclassified or misunderstood by the model, leading to inconsistent escalation. To address this, the AI team is tasked with making the model's decisions explainable using tools like LIME and SHAP. The goal is to interpret how input features (words/phrases) influence model predictions, compare local vs. global explanations, and identify cases of model bias or misclassification—thereby improving transparency and trustworthiness of the system. **Key Questions / Analysis / Interpretation to be evaluated during/after Implementation** Use LIME and SHAP to interpret predictions of a text classification model and analyze the influence of input features. 1. Train or load a pre-trained sentiment analysis model (e.g., Logistic Regression or BERT) on the IMDb or SST dataset. 2. Use LIME to generate local explanations for individual predictions by perturbing input tokens. 3. Apply SHAP to analyze global feature importance and visualize contribution scores across the dataset. 4. Compare LIME and SHAP outputs for the same predictions and interpret any discrepancies. **Supplementary Problems –** 1. How do LIME and SHAP explanations vary for complex or ambiguous sentences? 2. Apply explanations to incorrectly classified instances – what features misled the model? 3. Use explanations to identify and mitigate dataset bias (e.g., gendered language in reviews).	CO3, CO5/ PO6, PO7

	4. Visualize SHAP values over multiple inputs for a global model view.	
	Key Skills to be addressed – - Text preprocessing and model training - Understanding model explainability concepts - Hands-on usage of LIME and SHAP for NLP tasks - Interpretation and visualization of feature importance - Comparison between different explanation techniques **Applications –** Trustworthy AI Systems in healthcare, finance, and law **Learning Outcome –** • Understand the role of local vs. global explanations in model interpretability • Gain hands-on experience with LIME and SHAP to interpret NLP models • Learn to use interpretability tools for model • transparency and accountability	
	Dataset/Test Data (Source and Description If Applicable) : IMDb or SST-2 Dataset	
	Tools/Technology To Be Used: Python, Scikit-learn / Transformers (Hugging Face), LIME, SHAP, Pandas, Matplotlib,	
	*** Total Hours of Problem Definition Implementation** *** Total Hours of Engagement = Time required to implement solution + Modification time, if necessary + Testing of the solution by the faculty member (may be with suggested test data) :** 3 + 1 hours	
	Post Laboratory Work Description: Submit a **lab report** including: • Sample predictions and their LIME/SHAP explanations • Comparison charts and interpretation of key features • Reflections on usefulness and limitations of explainability techniques	

	Evaluation Strategy Including Viva: Model Training & Prediction Accuracy - 2 LIME/SHAP Implementation - 2 Visualization and Comparative Analysis- 2 Report Quality and Ethical Reflection - 2 Viva (Conceptual Understanding + Case - 2 Discussion) **Total 10 Marks** **Feedback on Problem Definition Implementation** **(Satisfaction Level 0 to 4, where 0 is lowest,1 is poor,2 is average, 3 is good, 4 is excellent) (This can asked for group of practical belongs to same tool/concept/technology):** **Satisfaction Level (0–4):** • 0: Not satisfied • 1: Poor • 2: Average • 3: Good • 4: Excellent **Feedback to be collected for:** • Ease of understanding the problem • Clarity of dataset and objectives • Relevance of tools/techniques • Difficulty level and learning impact	
6	**Problem Definition**: **Fair Classification and Bias Mitigation using Fairlearn Toolkit** A government agency is deploying an AI system to automate screening of applicants for financial aid. During testing, the model shows disproportionately lower approval rates for certain racial and gender groups. To ensure fairness and regulatory compliance, the data science team is assigned to audit the model's decisions and apply bias mitigation techniques using fairness toolkits. **Key Questions / Analysis / Interpretation to be evaluated during/after Implementation** Detect and mitigate fairness issues in a classification model predicting outcomes across different demographic groups.	CO3, CO5/ PO6, PO7

	1. Load the UCI Adult Income Dataset and split it into training/testing sets. 2. Train a baseline logistic regression or decision tree model to predict income class. 3. Analyze fairness metrics across sensitive attributes (e.g., gender, race) using Fairlearn's MetricFrame. 4. Apply bias mitigation strategies like reweighing or adversarial debiasing and compare results **Supplementary Problems –** 1. Evaluate the trade-off between fairness and accuracy when applying mitigation techniques. 2. Use different sensitive attributes (e.g., gender vs. race) and compare results. 3. Explore intersectional fairness (e.g., Black women vs. White men). 4. Implement pre-processing (reweighing) vs. post-processing (equalized odds post-processing) and compare. **Key Skills to be addressed –** - Model evaluation using fairness metrics - Application of constraint-based optimization for bias mitigation - Critical understanding of ethical trade-offs in ML - Data analysis and visualization for fairness evaluation - Practical experience with open-source fairness toolkits **Applications –** Fair hiring tools, Loan or credit scoring systems **Learning Outcome –** • Gain practical skills in measuring and diagnosing bias in AI systems • Learn how to apply post-processing or in- processing debiasing algorithms • Understand how fairness interventions affect model accuracy and equity	
	Dataset/Test Data (Source and Description If Applicable) : UCI Adult Dataset	
	Tools/Technology To Be Used: Python, Scikit-learn, Fairlearn, Pandas, Matplotlib/Seaborn	
	*** Total Hours of Problem Definition Implementation** *** Total Hours of Engagement = Time required to implement solution + Modification time, if necessary + Testing of the solution by the faculty member (may be with suggested test data) :** 3 + 1 hours	

	Post Laboratory Work Description: Submit a report including: • Confusion matrices, accuracy scores, and fairness metrics pre- and post-mitigation • Visualizations (e.g., bar charts for demographic parity) • Discussion of ethical trade-offs observed • Suggestions for further bias reduction techniques	
	Evaluation Strategy Including Viva: Model Training & Prediction Accuracy - 2 Fairness Evaluation using Fairlearn - 2 Bias Mitigation Implementation- 2 Visualizations and Ethical Discussion – 2 Viva (Conceptual Understanding + Case - Discussion) 2 **Total 10 Marks** **Feedback on Problem Definition Implementation (Satisfaction Level 0 to 4, where 0 is lowest,1 is poor,2 is average, 3 is good, 4 is excellent) (This can asked for group of practical belongs to same tool/concept/technology):** **Satisfaction Level (0–4):** • 0: Not satisfied • 1: Poor • 2: Average • 3: Good • 4: Excellent **Feedback to be collected for:** • Ease of understanding the problem • Clarity of dataset and objectives • Relevance of tools/techniques Difficulty level and learning impact	

7	**Problem Definition**: **Differential privacy limits the information that models can leak about individual training samples.** A health-tech startup is building an AI model to predict the risk of chronic diseases based on sensitive patient data like age, medical history, and lifestyle. While the model performs well, the company must ensure it complies with strict data privacy regulations such as HIPAA and GDPR. The concern is that attackers could extract information about individual patients from the trained model. To address this, the engineering team is assigned to implement Differential Privacy (DP) using the Opacus library in PyTorch, evaluate how much privacy is gained (ε), and understand the impact of privacy constraints on model performance. **Key Questions / Analysis / Interpretation to be evaluated during/after Implementation** 1. Train a classifier (e.g., on a health dataset) using Opacus (PyTorch) with differential privacy. 2. Compare accuracy and privacy loss (ε). **Supplementary Problems –** 1. Vary the noise multiplier and evaluate its effect on accuracy and privacy loss (ε). 2. Investigate model memorization by attempting membership inference (optional). 3. CombineDP with dropout or other regularization strategies. 4. Compare DP vs. data anonymization as privacy techniques. **Key Skills to be addressed –** -Understanding and implementation of differential privacy -Usage of DP-SGD in model training -Balancing privacy-utility trade-offs -Practical handling of privacy parameters (noise, clipping norm, ε-delta budget) - Evaluating models for privacy-aware deployment **Applications –** Financial risk scoring using sensitive user information, Federated learning with privacy constraints, Government and census data analytics **Learning Outcome –** • Implement privacy-preserving machine learning • Understand trade-offs between privacy and accuracy • Evaluate ε as a privacy metric	CO4, CO5/ PO7, PO8

Dataset/Test Data (Source and Description If Applicable) : MNIST (standard for private training demos)/Adult Census Dataset	
Tools/Technology To Be Used: PyTorch, Opacus	
*** Total Hours of Problem Definition Implementation** *** Total Hours of Engagement = Time required to implement solution + Modification time, if necessary + Testing of the solution by the faculty member (may be with suggested test data)** : 3 + 1 hours	
Post Laboratory Work Description: Submit a **report** including: • Accuracy and privacy loss values across experiments • Graphs showing the trade-off between model utility and privacy • Description of DP configuration parameters used • Reflections on practical implications and limitations of DP	
Evaluation Strategy Including Viva: 1. Model Training with and without DP - 2 2. Application and Tuning of DP Parameters - 2 3. Interpretation of Privacy-Utility Trade-off 2 4. Report Quality and Analytical Depth - 2 5. Viva (Conceptual Understanding + Case - Discussion) – 2 **Total 10 Marks** **Feedback on Problem Definition Implementation (Satisfaction Level 0 to 4, where 0 is lowest,1 is poor,2 is average, 3 is good, 4 is excellent) (This can asked for group of practical belongs to same tool/concept/technology):** **Satisfaction Level (0–4):** • 0: Not satisfied • 1: Poor • 2: Average • 3: Good • 4: Excellent **Feedback to be collected for:** • Ease of understanding the problem • Clarity of dataset and objectives • Relevance of tools/techniques • Difficulty level and learning impact	

8	**Problem Definition**: **Designing a Transparent and Inclusive Mental Health Chatbot** A healthcare startup is developing an AI-powered mental health assistant chatbot to support users experiencing symptoms of depression and anxiety. The assistant analyzes user conversations and offers emotional support and recommendations. However, due to the sensitive nature of mental health data, ethical risks such as lack of informed consent, misdiagnosis, data misuse, and exclusion of vulnerable groups must be addressed before deployment. To ensure responsible development, the startup's AI Ethics team is tasked with creating a prototype of the assistant that integrates informed consent, participatory feedback, and ethical safeguards through user-centered design. **Key Questions / Analysis / Interpretation to be evaluated during/after Implementation** Prototype and ethically evaluate a mental health chatbot using informed consent design, stakeholder feedback, and ethical impact analysis. 1. Load and explore the CLPsych 2015 dataset (Reddit posts labeled for mental health cues). 2. Simulate the chatbot's text classification pipeline (basic depression detection using Hugging Face Transformers or Naive Bayes). 3. Design a mock informed consent interface using Streamlit or Figma, explaining data usage and risks to users. 4. Collect mock feedback from stakeholders (students/faculty) via a Google Form on expectations, fears, and needs. 5. Complete an Ethics Impact Assessment (EIA) covering informed consent, potential harms, transparency, and participatory design insights. **Supplementary Problems –** 1. How would informed consent be handled in voice-based assistants vs. text-based? 2. Propose a feature to flag bias or misinformation in responses. 3. How can the assistant cater to low-literacy or differently-abled users? 4. Implement transparency logs showing how the assistant made its recommendations. **Key Skills to be addressed –** - Conversational AI and interface prototyping - Ethical design thinking and stakeholder engagement - Informed consent modeling and UI integration - Inclusive design practices and user experience analysis - Interdisciplinary integration of AI and healthcare regulations **Applications –** Fitness and lifestyle trackers with ethical safeguards, Mental health support assistants	CO4, CO5/ PO9, PO10

Learning Outcome –

- Apply ethical design principles (transparency, consent, participation) to a real-world use case
- Understand risks in sensitive applications like mental health AI
- Learn how to integrate stakeholder feedback into ethical AI design

Dataset/Test Data (Source and Description If Applicable) : CLPsych 2015 Reddit Dataset

- **Tools/Technology To Be Used:** Python, Transformers (Hugging Face), Streamlit or Figma, Google Forms, IRB-style Ethical Risk Template

*** Total Hours of Problem Definition Implementation**

*** Total Hours of Engagement = Time required to implement solution + Modification time, if necessary + Testing of the solution by the faculty member (may be with suggested test data)** : 3 + 1 hours

Post Laboratory Work Description:

Submit a report containing:

- Architecture of assistant and ethical modules
- Sample consent workflow and interaction screenshots
- Summary of participatory feedback and changes made
- Reflection on design choices and ethical implications

Evaluation Strategy Including Viva:

• Functional Assistant with Consent Logic - 2

• User Feedback Integration - 2

• Ethical Documentation and Design Rationale- 2

• UI/UX and Inclusivity Considerations – 2

• Viva (Conceptual Understanding + Case - Discussion) - 2

Total 10 Marks

Feedback on Problem Definition Implementation (Satisfaction Level 0 to 4, where 0 is lowest,1 is poor,2 is average, 3 is good, 4 is excellent) (This can asked for group of practical belongs to same tool/concept/technology):

Satisfaction Level (0–4):

- 0: Not satisfied
- 1: Poor
- 2: Average
- 3: Good
- 4: Excellent

	Feedback to be collected for: • Ease of understanding the problem • Clarity of dataset and objectives • Relevance of tools/techniques • Difficulty level and learning impact	
9	**Problem Definition**: **Auditing Bias in Algorithmic Risk Prediction Tools: The COMPAS Controversy** A state-level criminal justice department is using the COMPAS algorithm to assess the likelihood of reoffending among arrested individuals. The goal is to support judges in making bail and sentencing decisions. However, investigations by independent journalists and researchers have revealed that the system disproportionately assigns higher risk scores to Black defendants compared to White defendants, raising concerns of algorithmic bias and structural discrimination.To prepare for a public ethics review, the department commissions a data science team to analyze the fairness of the COMPAS system, propose mitigation strategies, and simulate a panel discussion on the ethical implications of deploying AI in the legal system. **Key Questions / Analysis / Interpretation to be evaluated during/after Implementation** • Analyze COMPAS dataset and its fairness metrics. • Debate the ethical implications in a mock panel. • Propose improvement strategies or alternatives. **Supplementary Problems –** • How does changing the decision threshold affect different demographic groups? • Evaluate intersectional fairness (e.g., young Black males vs. older white females). • Compare COMPAS with a simple rule-based heuristic model—does complexity ensure fairness? • Explore the impact of excluding race as a feature—does bias still persist? **Key Skills to be addressed –** • Model development and evaluation • Understanding of fairness metrics (statistical parity, false positive/negative rates) • Ethical reasoning in socio-technical systems • Hands-on experience with fairness toolkits • Ability to communicate fairness trade-offs clearly **Applications –** Employment or insurance screening tools, Credit scoring and financial models, Public policy and algorithmic decision-making	CO5, CO6/ PO10, PO12

	Learning Outcome – • Understand AI risks in the legal system • Evaluate fairness in decision-making systems • Develop policy awareness in high-stakes • Domains	
	Dataset/Test Data (Source and Description If Applicable) : COMPAS Dataset	
	Tools/Technology To Be Used: Python, AIF360, Google Docs (for collaborative panel prep)	
	*** Total Hours of Problem Definition Implementation** *** Total Hours of Engagement = Time required to implement solution + Modification time, if necessary + Testing of the solution by the** **faculty member (may be with suggested test data) :** 3 + 1 hours	
	Post Laboratory Work Description: Submit a report containing: • Pre- and post-mitigation performance and fairness metric values • Visualizations of disparities (e.g., bar plots of FPR/FNR across groups) • Summary of chosen fairness definitions and their interpretation • Reflection on how AI may reinforce or reduce structural biases	

	Evaluation Strategy Including Viva: • Data Preprocessing and Model Training - 2 • Fairness Evaluation with AIF360/Fairlearn - 2 • Mitigation Strategy and Re-evaluation- 2 • Ethical Reflection and Report Quality – 2 • Viva (Conceptual Understanding + Case - Discussion) - 2 **Total 10 Marks** **Feedback on Problem Definition Implementation (Satisfaction Level 0 to 4, where 0 is lowest,1 is poor,2 is average, 3 is good, 4 is excellent) (This can asked for group of practical belongs to same tool/concept/technology):** **Satisfaction Level (0–4):** • 0: Not satisfied • 1: Poor • 2: Average • 3: Good • 4: Excellent **Feedback to be collected for:** • Ease of understanding the problem • Clarity of dataset and objectives • Relevance of tools/techniques • Difficulty level and learning impact	
10	**Problem Definition**: **Interpreting Deep Learning Decisions in Critical AI Systems Using GradCAM** A medical AI company is deploying a deep learning model to detect pneumonia in chest X-rays**.** The model shows high accuracy, but doctors are reluctant to trust it because the predictions lack transparency. During validation, some misclassifications reveal that the model is focusing on irrelevant features like hospital tags, corners, or background noise instead of lung regions. To improve trust, transparency, and ethical safety, the internal AI audit team is assigned to apply GradCAM on the trained model to visualize its focus areas, assess alignment with domain knowledge, and recommend improvements for more responsible use of the model. **Key Questions / Analysis / Interpretation to be evaluated during/after Implementation** 1. Train or load a CNN (e.g., ResNet or VGG) for image classification (e.g., dogs vs. cats or pneumonia detection). 2. Apply GradCAM to generate heatmaps for selected predictions. 3. Evaluate whether the highlighted regions are semantically	CO6,CO7/ PO11, PO12

	meaningful (i.e., responsible regions). 4. Reflect on potential risks of mislocalization (e.g., focusing on background or watermark). **5.** Discuss how explainability supports fairness, safety, and debugging in AI systems. **Supplementary Problems –** 1. Visualize GradCAM outputs for both correct and incorrect predictions — is the model "looking" at the right regions? 2. Apply GradCAM to adversarial examples — how does focus shift? 3. Use GradCAM on biased datasets (e.g., images with confounding factors like hospital labels) and assess unintended focus. 4. Compare GradCAM with other interpretability methods (e.g., LIME or SHAP for vision). **Key Skills to be addressed –** - Deep learning model interpretation - Visual analytics and heatmap generation - Critical thinking on model behavior and decision-making - Connecting explainability with fairness, trust, and model debugging - Basic model training or inference using pretrained networks **Applications –** Medical imaging AI, AI auditing and compliance, Security and surveillance **Learning Outcome –** • Understand how GradCAM visualizes model decision rationale • Use interpretability tools to inspect CNN behavior • Identify and critique ethically risky or misaligned model focus - Advocate for responsible AI practices through explainability mechanisms	
	Dataset/Test Data (Source and Description If Applicable) : Cats vs. Dogs Dataset, Chest X-ray dataset (e.g., Pneumonia Detection) from Kaggle or NIH	
	Tools/Technology To Be Used: Python, TensorFlow, Torchvision, OpenCV, Matplotlib for visualization, GradCAM utility libraries (e.g., pytorch-gradcam)	

	*** Total Hours of Problem Definition Implementation** *** Total Hours of Engagement = Time required to implement solution + Modification time, if necessary + Testing of the solution by the faculty member (may be with suggested test data) :** 3 + 1 hours	
	Post Laboratory Work Description: Submit a report with: • Heatmap outputs for correct and incorrect classifications • Qualitative assessment of whether GradCAM output is "trustworthy" • Ethical interpretation	
	Evaluation Strategy Including Viva: Model Training or Loading - 2 GradCAM Heatmap Generation & Analysis - 2 Visualization and Ethical Interpretation - 2 Report Clarity and Justification - 2 Viva (Conceptual Understanding + Case - Discussion) 2 **Total 10 Marks** **Feedback on Problem Definition Implementation (Satisfaction Level 0 to 4, where 0 is lowest,1 is poor,2 is average, 3 is good, 4 is excellent) (This can asked for group of practical belongs to same tool/concept/technology):** **Satisfaction Level (0–4):** • 0: Not satisfied • 1: Poor • 2: Average • 3: Good • 4: Excellent **Feedback to be collected for:** • Ease of understanding the problem • Clarity of dataset and objectives • Relevance of tools/techniques • Difficulty level and learning impact	